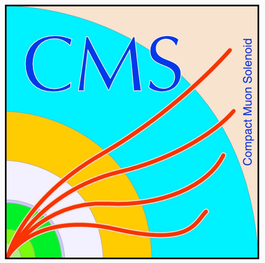
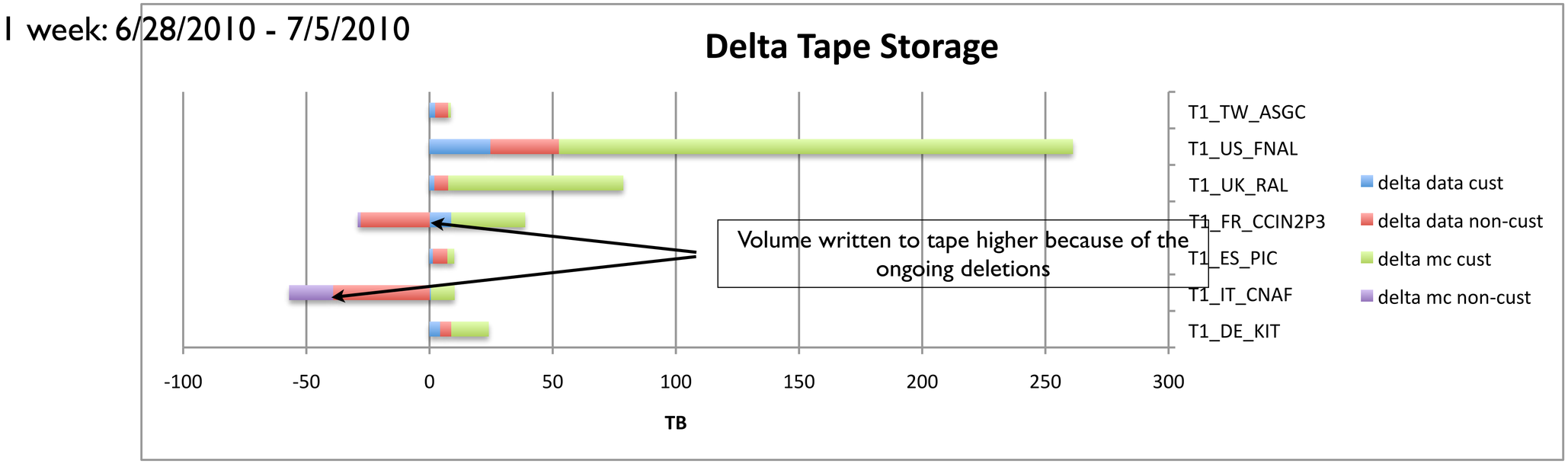
# Tier-1: Workflow Efficiency and IO

WLCG Collaboration Workshop 7-9 July 2010
CMS Experiment Jamboree




Oliver Gutsche
for
CMS Data Operations

# Introduction




- In preparation for ICHEP, CMS produced large volumes of data and MC and wrote them to tape at T1 sites
- We got reports from sites that at the current scale we are running into tape migration backlogs
  - These backlogs affect how fast and efficiently data from the detector can be archived and disturb the whole site operations
- In the following, we want to analyze how much data or different workflows are writing
- With this information and the site tape writing capabilities, we can determine new working points in terms of parallel jobs for sustained operation

# Workflows

- Primary Tier-1 workflows
  - Data re-reconstruction
  - Data skimming
  - MC re-digitization / re-reconstruction

**Data Re-reconstruction**

Input:
RAW input data dataset

Outputs (separate datasets):
RECO
ALCARECO (multiple)
DQM

**MC re-digitization / re-reconstruction**

Input:
RECO input data dataset

Skimmed Outputs (separate datasets, depends on configuration):
RECO
RAW-RECO
USER

**MC re-digitization / re-reconstruction**

Input:
GEN-SIM-RAW input MC dataset

Outputs (separate datasets):
GEN-SIM-RAW (new)
GEN-SIM-RECO
AODSIM
DQM

# Procedure

- Investigate how much data the different workflows produce per executable/core and per sec.

- Use success tarballs from the ProdAgent to extract properties from FrameworkJobReport per successful processing job
  - Sum of size of all output files
  - Job length (AppEndTime-AppStartTime)

- Important:
  - Script gives average of processing jobs writing into unmerged area
  - Everything has to be merged afterwards if not written directly to merged
    - This effect is neglected

# Results

Average output per executable [MB/s]

| | Output [MB/s] |
|---|---|
| ReReco | 0.180 |
| SD/CS skim | 0.400 |
| Redigi/rereco | 0.275 |

Updated 7/26/10 by fixing a bug

- Skimming writes out the most data volume per sec. per core
- Redigi/rereco writes out the largest data volumes
  - Need to discuss with Physics/Offline if writing out RAW is actually needed
    - Could reduce the output volume significantly
- But also other workflows are writing out data at a higher rate than naively expected

# Projections

Output projections

| Slots | Output in 1 hour [TB] | | | Output in 24 hour [TB] | | | Rate [MB/s] | | |
|---|---|---|---|---|---|---|---|---|---|
| | ReReco | SD/CS Skim | Redigi/rereco | ReReco | SD/CS Skim | Redigi/rereco | ReReco | SD/CS Skim | Redigi/rereco |
| 500 | 0.31 | 0.69 | 0.47 | 7.42 | 16.48 | 11.33 | 90.00 | 200.00 | 137.50 |
| 1000 | 0.62 | 1.37 | 0.94 | 14.83 | 32.96 | 22.66 | 180.00 | 400.00 | 275.00 |
| 2000 | 1.24 | 2.75 | 1.89 | 29.66 | 65.92 | 45.32 | 360.00 | 800.00 | 550.00 |
| 6800 | 4.20 | 9.34 | 6.42 | 100.85 | 224.12 | 154.08 | 1,224.00 | 2,720.00 | 1,870.00 |

- Numbers speak for themselves

Updated 7/26/10 by fixing a bug

- Todo for DataOps:
  - Use writing performance of tape systems at Tier-1 sites to determine working points
    - Running a single workflow alone
    - Combining different workflows
    - Taking into account necessary parallel tape activity
      - Archiving data from Tier-0, MC from T2 level
      - Reading back data/MC from tape